The F of findable – Searching for existing data

October 11, 2018

Researchers are increasingly asked to make their research data – where possible – available to others for further research in a way that the datasets are FAIR: findable, accessible, interoperable and re-usable. This policy focus implies that there are also still a lot of datasets – legacy data – that are not so easy to find or to re-use. If you are a researcher looking for existing data, which search strategies could you use?

AUTHORS

Mareike Boom
Research Data Steward

PROJECTS

Getting to the core of crimmigration

What's the use of existing data?

As the previous blog post in this series has shown, even in research projects where the collection of new data takes a central stage, the availability of existing data can be crucial for the success of such projects. Without a map or photograph of an area, it is a lot more difficult to decide where to go for your research as geologist. There might be many reasons why a researcher would need existing data or research documentation. This can range from contextual data, comparative data with different temporal or special coverage, or as basis of a main data-set to be analyzed with a new question. Using existing data also makes research possible where there are no budgets for the collection of new empirical data.

In the novel "Beyond Sleep"[1] - used as example in the previous blog posts - both the doctoral candidate and the supervisor are presented as unsure about how to approach the search for existing data. There, personal contacts and a referral approach are being used. There are numerous examples to be found about the role of existing data in research projects. If you are curious, maybe just ask about it at the next research presentation that you attend.

Where to "hunt" for data?

Sometimes it is difficult for (beginning) researchers to get started with looking for existing data. What kinds of strategies could researchers apply to find

existing data – the FAIR research data, (open) government data and the datasets that are not so easy to find? This blog post will give you some ideas and starting points, but by no means is meant to be exhaustive or replace discipline or method specific teaching in methods and systematic search.

At the end of research projects, researchers increasingly choose to archive the datasets that they have been using in a dedicated data archive (repository), such as DANS in the Netherlands. Together with their dataset they provide all the information necessary for the online catalog to search for specific datasets, those that are of interest for geologists or legal scholars and provide another researcher with enough details to determine if a dataset would be of interest for them. In most cases researchers can download or request a dataset via the catalog. In other cases the entry refers them to another organization. Gregory et. al. (2018) advise how to approach the search via such repositories, as with an increasing number of datasets, the search becomes a challenge.

Overviews of research data repositories and related sources

There are many archives or repositories that focus on sustainable long term preservation of research data. These organizations could be found on organizational, national or international level. Some are connected to a specific discipline or type of data; some are more interdisciplinary. For researchers this might mean that they need to consult various catalogs and databases.

How can researchers identify the repositories that might be relevant for them? The global Registry of Research Data Repositories re3data.org, funded by the German Research Foundation (DFG), allows to search or browse for repositories according to disciplines, countries or types of data, for example. This registry also refers to international surveys such as the European Social Survey (the ESS), that could be relevant for a socio-legal research project in Europe.

Various institutions have compiled discipline specific lists with resources for research data, which can be research data or data from other sources. One example for social sciences would be the Wissenschaftszentrum Berlin für Sozialforschung gGmbH. Researchers interested in datasets collected by researchers of Dutch universities, could use the NARCIS dataset search function to look for datasets of various disciplines. For discipline-specific databases/repositories, you could check out the resources list on the website of the Open Science MOOC that is being developed at the moment.

Before you get lost in your search and selection of online catalogs and resources, do not hesitate to get help from specialized librarians in your institution or experts from the data repository that holds the datasets that you are interested in.

Why do researchers chose a certain "data publishing route"? In its *Expert Tour Guide on Data Management* the CESSDA Training Working Group (2017) explains the advantages and disadvantages of the various options open to researchers.

Statistical data & open government data

Not all existing data that researcher are interested to use, will have been created or collected for the purpose of research. Also public sector information can contain a wealth of relevant data. There are many overviews of statistical offices worldwide, such as the list maintained by the United Nations Statistics Division

For public sector open data there is an overview of the European open data catalogs on the European Union Open Data Portal [2]. These portals are not primarily aiming at researchers, but also for a broader group of social and economic actors interested in the re-use.

Taking the example of legal research, the sites could refer you to judicial and court statistics, crime statistics or court information. Of course it can still happen – depending on the region - that the dataset that you are looking for has not been integrated in such portals and you would have to contact organizations directly.

In other cases, as Marc van Opijnen describes in his blog post about the question of increasing the number of publications of rulings in the Netherlands, the datasets could be huge and complete, but millions of unstructured documents prove very difficult to search. According to him even the structured and searchable half a million Dutch rulings still await to be used in serious big data research.

Other sources and search options

There are enough examples of unsuccessful searches by researchers for data for reuse as well as scholarly research to show that the findability of data decreases with time as well as the likelihood of finding correct contact details of the researchers; cf. e.g. Vines et al. (2014). Research in 2005 and 2011 showed that the existence of rules/policies with regard to research data and ethical principles did not always mean that these were (completely) followed by the authors (Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2011; Wicherts, Borsboom, Kats, & Molenaar, 2006) This should by no means discourage researchers to start looking for the dataset that is most relevant for their research. There are always several options on how to approach the search.

Via search tools for publications researchers can try to identify publications that are based on data that are potentially relevant for them. This might involve a change in the search strategies to also include topics or disciplines that might be considered as not relevant when reviewing literature for other purposes. In the best case the publication itself contains an identifier or web link to the dataset(s). If this is not the case, the information about the author and the method section might provide you with enough information about the dataset and where and how it might be obtained.

If it is not a quite recent dataset and it needs to be requested from the author(s), the information in the publication might only be a starting point. Many researchers change their affiliation during their career and contact details change. Next to general online search strategies, when 'hunting' for data, it is worth checking if the author – especially when the name is quite frequently used – manages his/her name with an author Identifier such as ORCID that links together affiliations, publications, education and so forth. Other types of online profiles also might list datasets as part of the researcher's publications.

If at this point no strategy has been successful, but the publication gives you the strong idea that the dataset used would be the key to the research, get advice from a colleague working within your discipline and/or librarian and map out the further options available depending on the information provided in the publication, e.g.:

- Look up citations of the publication/dataset and contact the authors of these more recent publications to see if they might have a copy of the dataset
- Contact the publisher
- Contact the institution the researcher worked for
- Contact a co-author
- Contact former colleagues of the author / members of the research group
- Contact other types of organizations that have been involved in the data collection
- ...

For sure there are more options, but this short blog post has hopefully provided basic strategies on how to look for existing (research) data.

Next steps

Did you find the dataset crucial for your research? Make sure to sort out the conditions for reuse (e.g. citations and credits) and any legal aspects before using it. If in doubt, contact relevant research support services at your institution.

This blog post is part of the series "Research in fiction through the lens of data management".

How to cite this blog post (Harvard style)

Boom, M.S. (2018) The F of findable – Searching for existing data. Available at: http://europeanbordercommuniti... (Accessed [date]).

Footnotes

[1] For international readers this blog post refers to an English translation Hermans, W.F. (2007). Beyond Sleep. (I. Rilke, Trans.). New York, NY: The Overlook Press. (Original work published 1966, translation of the 27th impression published in 2003 by De Bezige Bij).

[2] The website of the European Commission gives an overview of the implementation of the Directive on the re-use of public sector information (Directive 2003/98/EC).

Universiteit Leiden

Van Vollenhoven Institute for Law, Governance & Society

NWO Netherlands Organisation for Scientific Research

European Border Communities • Van Vollenhoven Institute • Leiden Law School, Universiteit Leiden • Steenschuur 25 • 2311 ES Leiden, The Netherlands • Disclaimer

Follow Van Vollenhoven Institute on Twitter